A balanced truncation based strategy for optimal control of evolution problems

Juan Carlos De Los Reyes[†] Tatjana Stykel[‡]

(December 11, 2009)

In this paper we present a balanced truncation based strategy for the numerical solution of optimal control problems governed by nonlinear evolution partial differential equations. The idea consists in utilizing a balanced truncation model reduction method for the efficient solution of the semi-discretized adjoint system, while the nonlinear state equations are fully solved. Our strategy is analyzed as a descent method in function spaces and global convergence results are presented. In combination with a Broyden-Fletcher-Goldfarb-Shanno update also superlinear convergence is verified. Numerical examples are given to illustrate the behaviour of the proposed method for different problems.

Keywords: Optimal control, adjoint equation, model reduction, balanced truncation

AMS Subject Classification: 49J15; 49J20; 49M05; 65K05; 93C15

1 Introduction

Model reduction techniques have been intensively investigated as a tool for the fast solution and optimization of evolution partial differential equations (PDEs). Among other methodologies, proper orthogonal decomposition (POD) and balanced truncation have been applied for the simulation of a wide range of phenomena including coherent structures, molecular dynamics and fluid flow, e.g., [1, 4, 16, 22, 35].

Proper orthogonal decomposition is currently the commonly used model reduction technique for nonlinear systems. The key idea of the POD method consists in choosing appropriate snapshots that are used to compute the reduced basis. A reduced-order model is then determined by Galerkin projection

[†] Departamento de Matemática, EPN Quito, Ecuador and Institut für Mathematik, TU Berlin, Germany, juan.delosreyes@epn.edu.ec. Supported by the DFG Sonderforschungsbereich 557 "*Control of complex turbulent shear flows*".

[‡] Institut für Mathematik, MA 4-5, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, stykel@math.tu-berlin.de. Supported by the DFG Research Center Matheon in Berlin.
Preprint 2009/28, Institut für Mathematik, Technische Universität Berlin, December 2009.

utilizing these basis functions. In the optimal control context, the application of this technique to the state equation and the cost functional yields a lower dimensional optimization problem which is usually easier to solve than the original one. Such a reduced-order problem is governed by ordinary differential equations and, consequently, numerical methods for the solution of finite dimensional optimal control problems or, after discretization in time, large-scale optimization routines are utilized. Note that the selection of the snapshots is highly problem-specific and has to be performed several times if optimization methods are involved, see e.g., [16, 10].

On the other hand, the balanced truncation model reduction approach has been previously proposed for linear time-invariant control systems [24, 7] and then extended to nonlinear systems in [19, 29, 32]. Balanced truncation methods for linear systems employ numerical linear algebra techniques and can be applied to large-scale problems [4, 5], whereas computational issues of balanced truncation algorithms for nonlinear systems remain a challenge. The connection between POD and balanced truncation is discussed in [30, 39].

In this paper, we propose an alternative approach for evaluating adjoints in the context of optimization methods for the solution of optimal control problems governed by nonlinear PDE. The idea consists in applying the balanced truncation method for the dimension reduction of the semi-discretized adjoint system, which is linear with respect to the adjoint variables. Since no reduction of the nonlinear state equation takes place, the model reduction error is present only in the adjoint system. In contrast to gradient evaluation techniques, see [8,9], the balanced truncation model reduction of the adjoint system makes it possible to obtain appropriate error bounds that allow a convergence analysis of the underlying optimization method. Here such an analysis is presented for descent methods including the Broyden-Fletcher-Goldfarb-Shanno method and leaving Newton-type methods for future work.

Note that the adjoint system contains time-varying linear terms coming from the nonlinearity in the state equation. To reduce the order of such a system we could apply the balanced truncation method adopted for the linear time-varying systems [31, 33, 38]. However, the time-varying terms depend on the state variable that changes at every iteration step of the optimization process. To avoid re-computing the reduced-order adjoint system for different state variables, we suggest to perform model reduction once by projecting the adjoint system onto a lower dimensional subspace. This subspace is determined using the balanced truncation method applied to the time-invariant system obtained from the adjoint system by dropping the state-dependent terms. Under some restrictions on the size of these terms and the model reduction error we can guarantee convergence of the descent method.

The efficiency of the balanced truncation model reduction method strongly depends on the size of the control parameters and the observation domain. In

this method, two matrix Lyapunov equations have to be solved. This may be quite expensive for large-scale problems if the number of inputs and outputs is large. However, for small number of inputs and outputs, the right-hand side of the Lyapunov equations has low rank. In this case there exist efficient iterative methods for solving such equations of very large state space dimensions, see [3, 21, 25].

The outline of the paper is as follows. In Section 2, the optimal control problem for two classes of nonlinear PDE is stated and the gradient of the cost functional for the infinite dimensional problem and the semi-discretized control problem are characterized. A general descent algorithm is also given and its convergence is analyzed. In Section 3, we briefly describe the balanced truncation model reduction approach and present the balanced truncation descent method for solving the semi-discretized optimal control problem. Sufficient conditions for the convergence of this method are investigated. Finally, in Section 4, numerical experiments are carried out in order to illustrate the efficiency of the proposed method.

2 Optimal control problem

Throughout the paper let $(\cdot\,,\cdot)_H$ denote the inner product and $\|\cdot\|_H$ the norm in a Hilbert space H. The topological dual of H is denoted by H' and the duality pairing is written as $\langle\cdot,\cdot\rangle_{H',H}$. Let H_1 and H_2 be two Hilbert spaces and let $\mathcal{L}(H_1, H_2)$ be the set of linear operators mapping H_1 into H_2. We will denote by $\mathcal{A}^\star$ the adjoint operator of $\mathcal{A} \in \mathcal{L}(H_1, H_2)$. The space of $n \times m$ real matrices is denoted by $\mathbb{R}^{n,m}$, the matrix $A^\top$ stands for the transpose of $A \in \mathbb{R}^{n,m}$ and $\|\cdot\|$ denotes the Euclidean vector norm or the spectral matrix norm.

2.1 *State equation*

Let V and H be two Hilbert spaces such that $V \hookrightarrow H \hookrightarrow V'$ with dense and continuous injections and let $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, be a regular bounded domain with a boundary Γ. We consider a general evolution equation

$$\begin{aligned} &\frac{\partial y}{\partial t} + \mathcal{A}\, y + \mathcal{N}(y) = f \qquad \text{in } \ (0,T)\times\Omega, \\ &y(0,x) = g(x) \qquad\qquad\qquad \text{in } \ \Omega, \end{aligned} \tag{1}$$

where $f \in L^2(0,T;H)$, $g \in H$ and $\mathcal{A} : V \to V'$ is a linear elliptic operator. Furthermore, $\mathcal{N} : V \to V'$ is a nonlinear operator given either by a polynomial

operator of the type

$$\mathcal{N}(y) = \sum_{j=0}^{2l-1} b_j y^j, \tag{2}$$

with positive integer l and positive b_{2l-1}, or by

$$\mathcal{N}(y) = \mathcal{N}_2(y, y), \tag{3}$$

where $\mathcal{N}_2(v, w)$ is a bilinear operator such that

$$\begin{aligned}
&(\mathcal{N}_2(v,w),w)_H = 0, && \forall v, w \in V,\\
&|(\mathcal{N}_2(v,w),z)_H| \le c_1 \|v\|_H^{\theta_1} \|v\|_V^{1-\theta_1} \|w\|_V \|z\|_V^{\theta_1} \|z\|_H^{1-\theta_1}, && \forall v, w, z \in V,\\
&\|\mathcal{N}_2(v,w)\|_H + \|\mathcal{N}_2(w,v)\|_H \le c_2 \|v\|_V \|w\|_V^{1-\theta_2} \|\mathcal{A}w\|_H^{\theta_2}, && \forall v \in V, w \in \mathcal{D}_{\mathcal{A}},\\
&\|\mathcal{N}_2(v,w)\|_H \le c_3 \|v\|_H^{\theta_3} \|v\|_V^{1-\theta_3} \|w\|_V^{1-\theta_3} \|\mathcal{A}w\|_H^{\theta_3}, && \forall v \in V,\ w \in \mathcal{D}_{\mathcal{A}}.
\end{aligned} \tag{4}$$

Here $\theta_i \in [0, 1)$ and c_i, $i = 1, 2, 3$, are positive constants, and $\mathcal{D}_{\mathcal{A}}$ denotes the domain of $\mathcal{A}$. Note that the boundary conditions in (1) are considered to be included in the definition of the spaces.

Example 2.1 Polynomial operators arises, for example, in the study of superconductivity of fluids, see [36, p. 98]. Specifically, we have the equations

$$\frac{\partial y}{\partial t} - D\Delta y = (1 - |y|^2)y$$

with appropriate initial and boundary conditions. Here D is a diagonal matrix with positive diagonal elements. The goal is to determine a function $y \in L^2(0, T; V)$ that satisfies this system on a bounded domain Ω.

Example 2.2 Bilinear operators satisfying (4) arise in fluid dynamics, see [36]. A classical example is given by the evolutionary Navier-Stokes equations

$$\begin{aligned}
&\frac{\partial y}{\partial t} - \nu\Delta y + (y \cdot \nabla)y + \nabla p = f,\\
&\text{div } y = 0,\\
&y|_\Gamma = 0, \quad y(0, x) = g(x),
\end{aligned}$$

where y is the velocity field, p is the pressure and ν is the viscosity coefficient of the fluid. The bilinear operator in this case is defined in the space $V = \{v \in (H_0^1(\Omega))^d : \text{div } v = 0\}$ as $\mathcal{N}_2(v, w) = (v \cdot \nabla)w$.

We consider the evolution equation (1) in the following weak form

$$\begin{aligned} &\frac{d}{dt}\langle y,\varphi\rangle_{V',V} + \langle \mathcal{A}\, y,\varphi\rangle_{V',V} + \langle \mathcal{N}(y),\varphi\rangle_{V',V} = \langle f,\varphi\rangle_{V',V}, \\ &y(0) = g, \end{aligned} \tag{5}$$

with $\varphi \in V$. A function $y \in L^2(0,T;V)$ is a weak solution of the state equation (1) if it satisfies (5) for all $\varphi \in V$ in a distributional sense on $(0,T)$. In the following theorem existence and uniqueness results for the state equation with nonlinear operators as in (2) and (4) are summarized.

THEOREM 2.3 *Consider* (5) *with a linear elliptic operator* $\mathcal{A}$ *and* $g \in H$.

(i) Let $V = H_0^1(\Omega), H = L^2(\Omega)$, $f \in L^2(0,T;H)$ *and let* $\mathcal{N}$ *be as in* (2). *There exists a unique solution* $y \in L^2(0,T;V) \cap L^{2l}(0,T;L^{2l}(\Omega))$ *for all* $T > 0$. *Moreover, if* $g \in V$, *then* $y \in C(0,T;V) \cap L^{2l}(0,T;H^2(\Omega))$ *for all* $T > 0$.

(ii) Let $\mathcal{N}(y) = \mathcal{N}_2(y,y)$ *with* $\mathcal{N}_2$ *satisfying* (4). *If* $f \in H$, *then there exists a unique solution* $y \in C([0,T];H) \cap L^2(0,T;V)$ *for all* $T > 0$. *Additionally, if* $g \in V$, *then* $y \in C([0,T];V) \cap L^2(0,T;\mathcal{D}_\mathcal{A})$ *for all* $T > 0$.

Proof The results are obtained by applying a Faedo-Galerkin technique. For the complete proofs we refer to [36] and [16], respectively. □

2.2 *Control problem*

Let U and Y be the Hilbert control and observation spaces, respectively. We consider the following abstract tracking type optimal control problem

$$\begin{aligned} \text{minimize} \quad & J(y,u) = \frac{1}{2}\int_0^T \|\mathcal{C}y - z\|_Y^2\, dt + \frac{\alpha}{2}\int_0^T \|u\|_U^2\, dt \\ \text{subject to} \quad & \frac{\partial y}{\partial t} + \mathcal{A}\, y + \mathcal{N}(y) = \mathcal{B}u \quad \text{in } (0,T)\times\Omega, \\ & y(0,x) = g(x) \qquad\qquad\quad \text{in } \Omega, \end{aligned} \tag{6}$$

where $\alpha > 0$, $g \in H$ and $z \in Y$ is a desired state. The operators $\mathcal{B} : U \to V'$ and $\mathcal{C} : V \to Y$ are the linear continuous control and observation operators, respectively. In practice, the control acts often on a subdomain of Ω only and not the whole state y is available for measurements. In this case the control and observation operators are just the extension and restriction operators, respectively. We will assume that the optimal control problem (6) admits an optimal solution, that is, there exists an optimal pair (y^*, u^*) that minimizes the cost functional J.

Let the control-to-state operator

$$\begin{aligned} \mathcal{G} : L^2(0,T;U) &\to L^2(0,T;V), \\ u &\mapsto y(u) \end{aligned}$$

be twice Fréchet differentiable. Moreover, we assume that its first derivative at the optimal solution $\mathcal{G}'(u^*)$ is a bijective linear operator. Then the first derivative $\hat{y} := \mathcal{G}'(u)v$ is characterized by the solution of the equation

$$\frac{\partial \hat{y}}{\partial t} + \mathcal{A}\hat{y} + \mathcal{N}'(y)\hat{y} = \mathcal{B}v, \quad \hat{y}(0) = 0, \tag{7}$$

where $\mathcal{N}'(y)$ is the Fréchet derivative of the operator $\mathcal{N}$ at y (see [37, Satz 5.10]). The cost functional in the optimal control problem (6) can, thus, be expressed in reduced form as

$$\mathcal{J}(u) := J(\mathcal{G}(u), u) = \frac{1}{2}\int_0^T \|\mathcal{C}\mathcal{G}(u) - z\|_Y^2 \, dt + \frac{\alpha}{2}\int_0^T \|u\|_U^2 \, dt. \tag{8}$$

A first-order optimality condition for $\min\limits_u \mathcal{J}(u)$ is then given by

$$\int_0^T (\mathcal{J}'(u^*), v)_U \, dt = 0 \quad \text{for all } \ v \in U, \tag{9}$$

where $\mathcal{J}'(u^*)$ stands for the Riesz representative of the Fréchet derivative of $\mathcal{J}$ at the optimal control u^*. To characterize this derivative, let us proceed formally from (8). Using the chain rule, we get that

$$\int_0^T (\mathcal{J}'(u), v)_U \, dt = \int_0^T \langle \mathcal{C}^\star(\mathcal{C}y - z), \hat{y}\rangle_{V',V} \, dt + \alpha \int_0^T (u, v)_U \, dt. \tag{10}$$

Introducing an adjoint state $p \in L^2(0,T;V)$ as the unique weak solution of

$$-\frac{\partial p}{\partial t} + \mathcal{A}^\star p + \mathcal{N}'(y)^\star p = -\mathcal{C}^\star(\mathcal{C}y - z), \quad p(T) = 0, \tag{11}$$

we have

$$\int_0^T (\mathcal{J}'(u), v)_U \, dt = \int_0^T \langle \frac{\partial p}{\partial t} - \mathcal{A}^\star p - \mathcal{N}'(y)^\star p, \hat{y} \rangle_{V',V} \, dt + \alpha \int_0^T (u, v)_U \, dt$$
$$= \int_0^T \langle \frac{\partial p}{\partial t}, \hat{y} \rangle_{V',V} \, dt - \int_0^T \langle \mathcal{A}\hat{y} + \mathcal{N}'(y)\hat{y}, p \rangle_{V',V} \, dt + \alpha \int_0^T (u, v)_U \, dt.$$

Using integration by parts in time [6, p. 477] we obtain that

$$\int_0^T (\mathcal{J}'(u), v)_U \; dt = (p(T), \hat{y}(T))_H - (p(0), \hat{y}(0))_H$$
$$- \int_0^T \langle \frac{\partial \hat{y}}{\partial t}, p \rangle_{V',V} \, dt - \int_0^T \langle \mathcal{A}\hat{y} + \mathcal{N}'(y)\hat{y}, p \rangle_{V',V} \, dt + \alpha \int_0^T (u, v)_U \, dt,$$

which, considering the initial condition for $\hat{y}$ in (7) and the final condition for p in (11), implies

$$\int_0^T (\mathcal{J}'(u), v)_U \, dt = - \int_0^T \langle \frac{\partial \hat{y}}{\partial t} + \mathcal{A}\hat{y} + \mathcal{N}'(y)\hat{y}, p \rangle_{V',V} \, dt + \alpha \int_0^T (u, v)_U \, dt.$$

Utilizing (7) we obtain that

$$\int_0^T (\mathcal{J}'(u), v)_U \, dt = - \int_0^T (\mathcal{B}v, p)_H \, dt + \alpha \int_0^T (u, v)_U \, dt$$

and, therefore,

$$\mathcal{J}'(u) = \alpha u - \mathcal{B}^\star p \quad \text{in} \quad U.$$

The characterization of the derivative of the cost functional plays a key role not only in the necessary optimality condition but also in the different optimization algorithms.

2.3 *Semi-discretized optimal control problem*

By utilizing a space discretization scheme such as a finite difference or a finite element method, we obtain from (6) the following large-scale optimal control

problem

$$\text{minimize} \quad \boldsymbol{J}(\boldsymbol{y},\boldsymbol{u}) = \frac{1}{2}\int_0^T (\boldsymbol{C}\boldsymbol{y}-\boldsymbol{z})^\top \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{y}-\boldsymbol{z})dt + \frac{\alpha}{2}\int_0^T \boldsymbol{u}^\top \boldsymbol{R}\boldsymbol{u}\,dt \quad (12)$$

$$\text{subject to} \quad \boldsymbol{E}\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) + \boldsymbol{B}\boldsymbol{u}, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0, \quad (13)$$

where $\boldsymbol{A} \in \mathbb{R}^{n,n}$ is a stiffness matrix, $\boldsymbol{B} \in \mathbb{R}^{n,m}$ is an input matrix, $\boldsymbol{C} \in \mathbb{R}^{q,n}$ is an output matrix, and matrices $\boldsymbol{E} \in \mathbb{R}^{n,n}$, $\boldsymbol{Q} \in \mathbb{R}^{q,q}$ and $\boldsymbol{R} \in \mathbb{R}^{m,m}$ are symmetric, positive definite. Additionally, $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{y}_0 \in \mathbb{R}^n$, $\boldsymbol{z} \in \mathbb{R}^q$ and $\boldsymbol{u} \in \mathbb{R}^m$ are the semi-discretized state, initial state, desired state and control, respectively, and $\dot{\boldsymbol{y}}$ denotes the time derivative of $\boldsymbol{y}$.

A function $\boldsymbol{y}$ is called a solution to (13) if it is absolutely continuous on $[0,T]$ and satisfies (13) in integral form

$$\boldsymbol{y}(t) = \boldsymbol{y}_0 + \int_0^t \left(\boldsymbol{E}^{-1}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{E}^{-1}\boldsymbol{N}(\boldsymbol{y}) + \boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{u}\right) d\tau. \quad (14)$$

The operator $\boldsymbol{N} : \mathbb{R}^n \to \mathbb{R}^n$ is a semi-discretized version of the nonlinear operator $-\mathcal{N}$ of type (2) or (3), (4). It has one of the following forms

$$\boldsymbol{N}(\boldsymbol{y}) = \sum_{j=0}^{2l-1} b_j \boldsymbol{y}^j \quad \text{or} \quad \boldsymbol{N}(\boldsymbol{y}) = diag(\boldsymbol{y})\mathbf{D}\boldsymbol{y},$$

where the power operation is to be understood componentwise and $\mathbf{D} \in \mathbb{R}^{n,n}$. The operator $\boldsymbol{N}$ is, therefore, twice differentiable. Consequently, when considered from $C([0,T])$ to $C([0,T])$, $\boldsymbol{N}$ constitutes a superposition operator, which is also twice Fréchet differentiable (cf. [18]). From the differentiability and the particular structure of the operator $\boldsymbol{N}$ under consideration, the following estimates follow

$$\left\|\boldsymbol{N}'(\boldsymbol{y}(t))\hat{\boldsymbol{y}}(t)\right\| \le \kappa_1(\boldsymbol{y}(t))\,\|\hat{\boldsymbol{y}}(t)\| \quad \text{for all } t \in [0,T], \quad (15)$$

$$\left\|\boldsymbol{N}''(\boldsymbol{y}(t))[\hat{\boldsymbol{y}}(t)]^2\right\| \le \kappa_2(\boldsymbol{y}(t))\,\|\hat{\boldsymbol{y}}(t)\|^2 \quad \text{for all } t \in [0,T], \quad (16)$$

with $\kappa_1(\boldsymbol{y}(t)) > 0$ and $\kappa_2(\boldsymbol{y}(t)) > 0$. In the case of the bilinear operator satisfying (4), we obtain that

$$\kappa_1(\boldsymbol{y}(t)) = K_{b,1}\|\boldsymbol{y}(t)\|, \qquad \kappa_2(\boldsymbol{y}(t)) = K_{b,2},$$

for some constants $K_{b,1} > 0$ and $K_{b,2} > 0$ that are independent of $\boldsymbol{y}$. For the

polynomial operator satisfying (2), we have

$$\kappa_1(\boldsymbol{y}(t)) \le \hat{K}_{p,1} + K_{p,1}\|\boldsymbol{y}(t)\|^{2l-2}, \qquad \kappa_2(\boldsymbol{y}) \le \hat{K}_{p,2} + K_{p,2}\|\boldsymbol{y}(t)\|^{2l-3}$$

for some constants $\hat{K}_{p,i},\ K_{p,i} > 0,\ i = 1, 2$, that are also independent of $\boldsymbol{y}$. In general, for both types of operators, there exist constants $\gamma_i > 0$ and $\hat{K}_i,\ K_i \ge 0$ such that

$$\begin{aligned} \kappa_1(\boldsymbol{y}(t)) &\le \hat{K}_1 + K_1\|\boldsymbol{y}\|^{\gamma_1}_{L^\infty(0,T)} =: \hat{\kappa}_1(\boldsymbol{y}), \\ \kappa_2(\boldsymbol{y}(t)) &\le \hat{K}_2 + K_2\|\boldsymbol{y}\|^{\gamma_2}_{L^\infty(0,T)} =: \hat{\kappa}_2(\boldsymbol{y}). \end{aligned} \tag{17}$$

Moreover, the following L^q-estimates follow from the differentiability properties

$$\left\|\boldsymbol{N}'(\boldsymbol{y})\hat{\boldsymbol{y}}\right\|_{L^q(0,T)} \le \sqrt[q]{T}\,\|\boldsymbol{N}'(\boldsymbol{y})\hat{\boldsymbol{y}}\|_{L^\infty(0,T)} \le \sqrt[q]{T}\hat{\kappa}_1(\boldsymbol{y})\,\|\hat{\boldsymbol{y}}\|_{L^\infty(0,T)}, \tag{18}$$

$$\left\|\boldsymbol{N}''(\boldsymbol{y})[\hat{\boldsymbol{y}}]^2\right\|_{L^q(0,T)} \le \sqrt[q]{T}\left\|\boldsymbol{N}''(\boldsymbol{y})[\hat{\boldsymbol{y}}]^2\right\|_{L^\infty(0,T)} \le \sqrt[q]{T}\hat{\kappa}_2(\boldsymbol{y})\,\|\hat{\boldsymbol{y}}\|^2_{L^\infty(0,T)}, \tag{19}$$

with $1 \le q < \infty$. The following theorem gives estimates for the solution of the initial problem (13).

THEOREM 2.4 *There exists a unique solution of the initial problem* (13) *on the interval* $[0, T]$. *This solution satisfies the following estimates*

$$\|\boldsymbol{y}\|_{L^q(0,T)} \le \sqrt[q]{T}\left(\|\boldsymbol{y}_0\| + \sqrt{T}\|\boldsymbol{E}^{-1}\|\,\|\boldsymbol{B}\|\,\|\boldsymbol{u}\|_{L^2(0,T)}\right) e^{\|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\|+\kappa_0)T}, \tag{20}$$

$$\|\boldsymbol{y}\|_{L^\infty(0,T)} \le \left(\|\boldsymbol{y}_0\| + \sqrt{T}\|\boldsymbol{E}^{-1}\|\,\|\boldsymbol{B}\|\,\|\boldsymbol{u}\|_{L^2(0,T)}\right) e^{\|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\|+\kappa_0)T}, \tag{21}$$

where $\kappa_0 = \max\limits_{0\le t\le T} \|\int_0^1 \boldsymbol{N}'(\tau\boldsymbol{y}(t))\,d\tau\|$ *and* $1 \le q < \infty$.

Proof Since $\boldsymbol{N}$ is differentiable, system (13) is solvable, see [2, p. 178]. Moreover, from the mean value theorem it follows that

$$\|\boldsymbol{N}(\boldsymbol{y}(t))\| \le \left\|\int_0^1 \boldsymbol{N}'(\tau\boldsymbol{y}(t))\,d\tau\right\| \|\boldsymbol{y}(t)\| \le \kappa_0\|\boldsymbol{y}(t)\|, \tag{22}$$

which, together with (14) implies that

$$\|\boldsymbol{y}(t)\| \le \|\boldsymbol{y}_0\| + \|\boldsymbol{E}^{-1}\|\,\|\boldsymbol{B}\|\int_0^t \|\boldsymbol{u}\|\,d\tau + \|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\| + \kappa_0)\int_0^t \|\boldsymbol{y}\|\,d\tau.$$

Utilizing Gronwall's inequality it follows that

$$\|\boldsymbol{y}(t)\| \leq (\|\boldsymbol{y}_0\| + \|\boldsymbol{E}^{-1}\|\,\|\boldsymbol{B}\|\|\boldsymbol{u}\|_{L^1(0,T)})e^{\|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\|+\kappa_0)T}$$

which by Hölder's inequality implies (20) and (21). □

We rewrite now the semi-discretized cost functional in reduced form as

$$\boldsymbol{\mathcal{J}}(\boldsymbol{u}) := \boldsymbol{J}(\boldsymbol{G}(\boldsymbol{u}),\boldsymbol{u}) = \frac{1}{2}\int_0^T (\boldsymbol{C}\boldsymbol{G}(\boldsymbol{u})-\boldsymbol{z})^\top \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{G}(\boldsymbol{u})-\boldsymbol{z})dt + \frac{\alpha}{2}\int_0^T \boldsymbol{u}^\top \boldsymbol{R}\boldsymbol{u}\,dt,$$

where $\boldsymbol{G} : L^2(0,T) \mapsto L^2(0,T)$ denotes the semi-discretized control-to-state operator. Using the implicit function theorem, the differentiability of the semi-discretized control-to-state mapping follows from the differentiability of the superposition operator $\boldsymbol{N}$.

Next, an estimate for the linearized state will be obtained. This estimate will be used thereafter in the convergence analysis of the proposed optimization method.

LEMMA 2.5 *The solution $\hat{\boldsymbol{y}}$ of the linearized equation*

$$\boldsymbol{E}\dot{\hat{\boldsymbol{y}}} = \boldsymbol{A}\hat{\boldsymbol{y}} + \boldsymbol{N}'(\boldsymbol{y})\hat{\boldsymbol{y}} + \boldsymbol{B}\boldsymbol{v}, \quad \hat{\boldsymbol{y}}(0) = 0 \tag{23}$$

satisfies the estimate

$$\|\hat{\boldsymbol{y}}\|_{L^\infty(0,T)} \leq \sqrt{T}e^{\kappa(\boldsymbol{y})}\,\|\boldsymbol{E}^{-1}\|\,\|\boldsymbol{B}\|\,\|\boldsymbol{v}\|_{L^2(0,T)} \tag{24}$$

with $\kappa(\boldsymbol{y}) = T\,\|\boldsymbol{E}^{-1}\|\,(\|\boldsymbol{A}\| + \hat{\kappa}_1(\boldsymbol{y}))$.

Proof Rewriting equation (23) in integral form, we get that

$$\hat{\boldsymbol{y}}(t) = \int_0^t \boldsymbol{E}^{-1}(\boldsymbol{A} + \boldsymbol{N}'(\boldsymbol{y}(\tau)))\hat{\boldsymbol{y}}(\tau)\,d\tau + \int_0^t \boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{v}(\tau)\,d\tau,$$

which by taking norms on both sides yields

$$\|\hat{\boldsymbol{y}}(t)\| \leq \|\boldsymbol{E}^{-1}\|\left(\int_0^t \|\boldsymbol{A} + \boldsymbol{N}'(\boldsymbol{y}(\tau))\|\|\hat{\boldsymbol{y}}(\tau)\|\,d\tau + \|\boldsymbol{B}\|\int_0^t \|\boldsymbol{v}(\tau)\|\,d\tau\right).$$

Gronwall's inequality implies that

$$\|\hat{\boldsymbol{y}}(t)\| \leq \|\boldsymbol{E}^{-1}\|\|\boldsymbol{B}\|e^{\|\boldsymbol{E}^{-1}\|\int_0^T \|\boldsymbol{A}+\boldsymbol{N}'(\boldsymbol{y}(\tau))\|d\tau}\int_0^T \|\boldsymbol{v}(\tau)\|\,d\tau. \tag{25}$$

From estimates (15) and (17) it follows that $\|\boldsymbol{A} + \boldsymbol{N}'(\boldsymbol{y}(\tau))\| \leq \|\boldsymbol{A}\| + \hat{\kappa}_1(\boldsymbol{y})$. Consequently, introducing $\kappa(\boldsymbol{y}) = T\ \|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\| + \hat{\kappa}_1(\boldsymbol{y}))$ and using Hölder's inequality, we obtain estimate (24). □

Taking into account the differentiability of the semi-discretized control- to-state mapping and proceeding formally as in Section 2.2, the derivative of $\boldsymbol{\mathcal{J}}(\boldsymbol{u})$ can be computed as

$$\boldsymbol{\mathcal{J}}'(\boldsymbol{u}) = \alpha \boldsymbol{R}\boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{p}, \tag{26}$$

where $\boldsymbol{p}$ is a solution of the adjoint equation

$$-\boldsymbol{E}\dot{\boldsymbol{p}} = \boldsymbol{A}^\top \boldsymbol{p} + \boldsymbol{F}(\boldsymbol{y})\, \boldsymbol{p} + \boldsymbol{C}^\top \boldsymbol{Q}(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{y}), \qquad \boldsymbol{p}(T) = 0, \tag{27}$$

with $\boldsymbol{F}(\boldsymbol{y}) = \boldsymbol{N}'(\boldsymbol{y})^\top$.

2.4 *Descent methods*

As a preparatory step for the convergence analysis of descent methods, we will begin by studying conditions for uniform continuity of the second derivative of the cost functional on the closed convex hull of the set

$$U_0 = \{\ \boldsymbol{u} \in L^2(0,T)\ :\ \boldsymbol{\mathcal{J}}(\boldsymbol{u}) \leq \boldsymbol{\mathcal{J}}(\boldsymbol{u}_0)\ \},$$

where $\boldsymbol{u}_0$ denotes the initial value for the optimization variable. The closed convex hull of U_0 will be denoted by $U_{\boldsymbol{\mathcal{J}}}$.

Proposition 2.6 *Let* $\boldsymbol{N} : C([0,T]) \to C([0,T])$ *be a twice differentiable operator such that* (18), (19) *hold. Then there exists a constant* $M > 0$ *such that*

$$\boldsymbol{\mathcal{J}}''(\boldsymbol{u})[\boldsymbol{v}]^2 \leq M\, \|\boldsymbol{v}\|^2_{L^2(0,T)} \tag{28}$$

for all $\boldsymbol{u} \in U_{\boldsymbol{\mathcal{J}}}$ *and* $\boldsymbol{v} \in L^2(0,T)$.

Proof Let $\boldsymbol{u} \in U_{\boldsymbol{\mathcal{J}}}$. It can be verified that the second derivative of the cost functional is given by

$$\boldsymbol{\mathcal{J}}''(\boldsymbol{u})[\boldsymbol{v}]^2 = \int_0^T \hat{\boldsymbol{y}}^\top \boldsymbol{C}^\top \boldsymbol{Q}\boldsymbol{C}\hat{\boldsymbol{y}}\, dt + \int_0^T \tilde{\boldsymbol{y}}^\top \boldsymbol{C}^\top \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}) dt + \alpha \int_0^T \boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{v}\, dt, \tag{29}$$

where $\tilde{\boldsymbol{y}}$ is solution of the equation

$$\boldsymbol{E}\dot{\tilde{\boldsymbol{y}}} = \boldsymbol{A}\tilde{\boldsymbol{y}} + \boldsymbol{N}'(\boldsymbol{y})\tilde{\boldsymbol{y}} - \boldsymbol{N}''(\boldsymbol{y})[\hat{\boldsymbol{y}}]^2, \quad \tilde{\boldsymbol{y}}(0) = 0. \tag{30}$$

Proceeding as in the proof of Lemma 2.5, we obtain from (19) the estimate

$$\begin{aligned} \|\tilde{\boldsymbol{y}}(t)\| &\le \sqrt{T} e^{\kappa(\boldsymbol{y})} \left\|\boldsymbol{E}^{-1}\right\| \left\|\boldsymbol{N}''(\boldsymbol{y})[\hat{\boldsymbol{y}}]^2\right\|_{L^2(0,T)} \\ &\le T e^{\kappa(\boldsymbol{y})} \left\|\boldsymbol{E}^{-1}\right\| \hat{\kappa}_2(\boldsymbol{y}) \|\hat{\boldsymbol{y}}\|^2_{L^\infty(0,T)}. \end{aligned} \tag{31}$$

Using Hölder and Cauchy-Schwarz inequalities we obtain that

$$\int_0^T \tilde{\boldsymbol{y}}^\top \boldsymbol{C}^\top \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z})\, dt \le \frac{1}{2}\|\boldsymbol{C}\| \, \|\boldsymbol{Q}\| \, \|\tilde{\boldsymbol{y}}\|_{L^\infty(0,T)} \left(T + \|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|^2_{L^2(0,T)}\right),$$

which, thanks to estimate (31) and since $\boldsymbol{u} \in U_{\mathcal{J}}$, yields that

$$\int_0^T \tilde{\boldsymbol{y}}^\top \boldsymbol{C}^\top \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z})\, dt \le c\, e^{\kappa(\boldsymbol{y})} \hat{\kappa}_2(\boldsymbol{y}) \, \|\hat{\boldsymbol{y}}\|^2_{L^\infty(0,T)}, \tag{32}$$

where $c = T\left(\mathcal{J}(\boldsymbol{u}_0) + \frac{T}{2}\right) \|\boldsymbol{C}\| \, \|\boldsymbol{Q}\| \, \|\boldsymbol{E}^{-1}\|$.

As a consequence of estimate (21), there exist constants $\hat{\theta}_i,\ \theta_i > 0$, $i = 1, 2$, such that

$$\kappa(\boldsymbol{y}) \le \hat{\theta}_1 + \theta_1 \|\boldsymbol{u}\|^{\gamma_1}_{L^2(0,T)}, \qquad \hat{\kappa}_2(\boldsymbol{y}) \le \hat{\theta}_2 + \theta_2 \|\boldsymbol{u}\|^{\gamma_2}_{L^2(0,T)}, \tag{33}$$

which, thanks to Cauchy-Schwarz inequality and since $\boldsymbol{u} \in U_{\mathcal{J}}$, implies that

$$\kappa(\boldsymbol{y}) \le \hat{\varpi}_1 + \varpi_1 \mathcal{J}(\boldsymbol{u}_0)^{\gamma_1}, \qquad \hat{\kappa}_2(\boldsymbol{y}) \le \hat{\varpi}_2 + \varpi_2 \mathcal{J}(\boldsymbol{u}_0)^{\gamma_2}, \tag{34}$$

with $\hat{\varpi}_i := \hat{\theta}_i + \theta_i^2/2$ and $\varpi_i := 2^{\gamma_i - 1} \|R^{-1/2}\|^{\gamma_i} / \alpha^{\gamma_i},\ i = 1, 2$.

Consequently, from (29), (32) and (34), we obtain that

$$\begin{aligned} \mathcal{J}''(\boldsymbol{u})[\boldsymbol{v}]^2 \le{}& \int_0^T \|\boldsymbol{Q}\| \, \|\boldsymbol{C}\|^2 \|\hat{\boldsymbol{y}}\|^2 dt + \alpha \int_0^T \|\boldsymbol{R}\| \, \|\boldsymbol{v}\|^2 dt \\ &+ c\left[\hat{\varpi}_2 + \varpi_2 \mathcal{J}(\boldsymbol{u}_0)^{\gamma_2}\right] \exp\left(\hat{\varpi}_1 + \varpi_1 \mathcal{J}(\boldsymbol{u}_0)^{\gamma_1}\right) \|\hat{\boldsymbol{y}}\|^2_{L^\infty(0,T)}, \end{aligned} \tag{35}$$

which, thanks to (24), implies the existence of a constant $M > 0$ such that (28) is fulfilled. □

Using (26) we can state the descent algorithm for the semi-discretized optimal control problem (12), (13). Following the general framework of descent methods in Hilbert spaces (see, e.g., [15,23]) and recalling that M is the bound on $\boldsymbol{\mathcal{J}}''(\boldsymbol{u})$ given by (28), we have the following algorithm.

Algorithm 2.7 Descent method

1. Consider an initial control $\boldsymbol{u}_0$.
2. FOR $k = 0, 1, \ldots$
   (a) find a direction $\boldsymbol{d}_k \neq 0$ such that

$$-(\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k), \boldsymbol{d}_k)_{L^2(0,T)} \geq \mu \|\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)\|_{L^2(0,T)} \|\boldsymbol{d}_k\|_{L^2(0,T)} \tag{36}$$

   with $\mu > 0$ and $\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k) \neq 0$;
   (b) set a line search step β_k such that

$$\boldsymbol{\mathcal{J}}(\boldsymbol{u}_k + \beta_k \boldsymbol{d}_k) < \boldsymbol{\mathcal{J}}(\boldsymbol{u}_k); \tag{37}$$

   (c) update $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \beta_k \boldsymbol{d}_k$.
   END FOR

Remark 2.8 In order to obtain convergence of Algorithm 2.7 a stronger condition on the step sizes is required. A classical condition is given by:

$$\boldsymbol{\mathcal{J}}(\boldsymbol{u}_k + \beta_k \boldsymbol{d}_k) \leq \boldsymbol{\mathcal{J}}(\boldsymbol{u}_k) - \frac{\mu^2}{2M} \|\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)\|, \tag{38}$$

where $\mu > 0$ is the same as in (36) and M is the constant from (28). Alternative conditions on the step sizes are given in, e.g., [11, Section 2.2]

THEOREM 2.9 *If the hypotheses of Proposition 2.6 hold and conditions* (36), (38) *are fulfilled at each iteration, then the sequence* $\boldsymbol{u}_k$ *generated by Algorithm 2.7 satisfies* $\lim_{k\to\infty} \boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k) = 0$. *Moreover, if there exists* $\varsigma > 0$ *such that*

$$\varsigma \|\boldsymbol{v}\|^2 \leq \boldsymbol{\mathcal{J}}''(\boldsymbol{u})[\boldsymbol{v}]^2$$

holds for all $\boldsymbol{u} \in U_{\boldsymbol{\mathcal{J}}}$ *and* $\boldsymbol{v} \in \mathbb{R}^m$, *then there exists* $\boldsymbol{u}^* \in U_{\boldsymbol{\mathcal{J}}}$ *such that* $\lim_{k\to\infty} \boldsymbol{u}_k = \boldsymbol{u}^*$ *and* $\boldsymbol{u}^*$ *is a solution of* (12), (13).

Proof Since (28) holds, the result follows from the general convergence theorem for descent methods in Hilbert spaces [23, pg. 288]. □

If the negative gradient is used as descent direction, i.e., $\boldsymbol{d}_k = -\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)$, condition (36) is immediately satisfied for $\mu = 1$. More generally, the directions of

the type $\boldsymbol{d}_k = -H_k \boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)$ with positive definite H_k satisfy (36). A particular example is given by the inverse of the operator

$$B_{k+1} = B_k + \frac{(\boldsymbol{z}_k, \cdot)_{L^2(0,T)}}{(\boldsymbol{d}_k, \boldsymbol{z}_k)_{L^2(0,T)}} \boldsymbol{z}_k - \frac{(B_k \boldsymbol{d}_k, \cdot)_{L^2(0,T)}}{(\boldsymbol{d}_k, B_k \boldsymbol{d}_k)_{L^2(0,T)}} B_k \boldsymbol{d}_k, \tag{39}$$

where $\boldsymbol{z}_k = \boldsymbol{\mathcal{J}}'(\boldsymbol{u}_{k+1}) - \boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)$. This operator update corresponds to the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. For this method a local superlinear convergence result additionally holds, see [14].

THEOREM 2.10 *Let $\boldsymbol{\mathcal{J}}$ be twice Fréchet differentiable and $\boldsymbol{\mathcal{J}}''(\cdot)$ Lipschitz continuous in a neighborhood of $\boldsymbol{u}^*$ with $\boldsymbol{\mathcal{J}}'(\boldsymbol{u}^*) = 0$ and bounded $\boldsymbol{\mathcal{J}}''(\boldsymbol{u}^*)^{-1}$. Let B_0 be a initial matrix iterate. If $B_0 - \boldsymbol{\mathcal{J}}''(\boldsymbol{u}^*)$ is compact, then the BFGS iterates converge q-superlinearly to $\boldsymbol{u}^*$ provided $\|\boldsymbol{u}_0 - \boldsymbol{u}^*\|_{L^2(0,T)}$ and $\|B_0 - \boldsymbol{\mathcal{J}}''(\boldsymbol{u}^*)\|_{\mathcal{L}(L^2(0,T)^2,\mathbb{R})}$ are sufficiently small.*

Remark 2.11 For a survey on descent methods for finite dimensional control problems, we refer to [27]. For the extension of second order optimization methods to optimal control problems, see, e.g. [13, 14, 17].

3 Balanced truncation descent method

From Algorithm 2.7 and equations (26), (27) it can be observed that the evaluation of the gradient of the cost functional in the descent method requires the numerical solution of the semi-discretized state and adjoint systems at each iterative step. The dimension of these systems depends on the level of refinement of the space discretization and is usually very large. Despite the ever increasing computational speed, the numerical solution of very large instationary problems is still a computationally intensive task especially when such problems have to be solved for different input functions. In this case the application of model order reduction can significantly reduce the computing time and memory requirements.

A general idea of model reduction is to approximate a large-scale system by a reduced model of lower dimension that has nearly the same behaviour as the original system. Although model reduction of nonlinear systems received a lot of attention in the last years, see [19,29,32], the development of efficient model reduction methods for such systems remains challenging from the mathematical and algorithmic points of view. Therefore, in this paper we propose to reduce the dimension of the linear adjoint system only, whereas the nonlinear state equation is fully solved. Even in this case the amount of computations can be reduced considerably.

The model order reduction problem for the semi-discretized adjoint equation

consists in an approximation of (27) by a reduced-order system of the same form

$$\begin{aligned} -\tilde{\boldsymbol{E}}\dot{\tilde{\boldsymbol{p}}} &= \big(\tilde{\boldsymbol{A}}^\top + \tilde{\boldsymbol{F}}(\boldsymbol{y})\big)\tilde{\boldsymbol{p}} + \tilde{\boldsymbol{C}}^\top \boldsymbol{Q}(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{y}), \quad \tilde{\boldsymbol{p}}(T) = 0, \\ \tilde{\boldsymbol{w}} &= \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}, \end{aligned} \tag{40}$$

with $\tilde{\boldsymbol{E}}$, $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{F}}(\boldsymbol{y}) \in \mathbb{R}^{\ell,\ell}$, $\tilde{\boldsymbol{B}} \in \mathbb{R}^{\ell,m}$, $\tilde{\boldsymbol{C}} \in \mathbb{R}^{q,\ell}$ and $\ell \ll n$. Since for computing the gradient $\mathcal{J}'(\boldsymbol{u})$ in (26) we need $\boldsymbol{B}^\top \boldsymbol{p}$ and not the whole vector $\boldsymbol{p}$, the utilization of the reduced-order model (40) is justified for approximating the output function $\boldsymbol{B}^\top \boldsymbol{p}$. In other words, we determine an approximate output $\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}$ such that the difference $\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}} - \boldsymbol{B}^\top \boldsymbol{p}$ is sufficiently small in some norm.

The approximate system (40) can be computed by a projection

$$\begin{gathered} \tilde{\boldsymbol{E}} = \boldsymbol{W}^\top \boldsymbol{E}\boldsymbol{S}, \qquad \tilde{\boldsymbol{A}} = \boldsymbol{S}^\top \boldsymbol{A}\boldsymbol{W}, \qquad \tilde{\boldsymbol{F}}(\boldsymbol{y}) = \boldsymbol{W}^\top \boldsymbol{F}(\boldsymbol{y})\boldsymbol{S}, \\ \tilde{\boldsymbol{B}} = \boldsymbol{S}^\top \boldsymbol{B}, \qquad \tilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{W}, \end{gathered} \tag{41}$$

where the projection matrices $\boldsymbol{W}$, $\boldsymbol{S} \in \mathbb{R}^{n,\ell}$ are, in general, time-varying. Note that the matrix $\boldsymbol{F}$ depends on the state vector $\boldsymbol{y}$ that changes at every iteration step in the descent algorithm. To avoid the re-computation of the reduced-order system that may be more expensive than solving one adjoint equation, we suggest to use the same time-invariant projection matrices $\boldsymbol{W}$ and $\boldsymbol{S}$ for all iterations. These matrices can be computed by a balanced truncation model reduction method [7, 20, 24] applied to system (27) with t replaced by $T - t$ and $\boldsymbol{F} = 0$. In the next subsection we briefly describe the basic idea of this method, see [7, 24] for more details.

3.1 *Balanced truncation*

Consider a linear time-invariant dynamical system

$$\begin{aligned} \boldsymbol{E}\,\dot{\xi}(t) &= \boldsymbol{A}^\top \xi(t) + \boldsymbol{C}^\top \eta(t), \quad \xi(0) = 0, \\ \varrho(t) &= \boldsymbol{B}^\top \xi(t). \end{aligned} \tag{42}$$

where $\boldsymbol{E}$ is symmetric and nonsingular. In the frequency domain this system can be rewritten as $\boldsymbol{\varrho}(s) = \boldsymbol{\mathcal{T}}(s)\boldsymbol{\eta}(s)$, where $\boldsymbol{\eta}(s)$ and $\boldsymbol{\varrho}(s)$ are the Laplace transforms of $\eta(t)$ and $\varrho(t)$, respectively, and $\boldsymbol{\mathcal{T}}(s) = \boldsymbol{B}^\top (s\boldsymbol{E} - \boldsymbol{A}^\top)^{-1}\boldsymbol{C}^\top$ is a *transfer function* of system (42). Balanced truncation model reduction is strongly related to the controllability Gramian $\mathcal{P}$ and the observability

Gramian $\mathcal{Q}$ of (42) which solve the generalized Lyapunov equations

$$\begin{aligned} \boldsymbol{E}\,\mathcal{P}\,\boldsymbol{A} + \boldsymbol{A}^\top\mathcal{P}\,\boldsymbol{E} &= -\boldsymbol{C}^\top\boldsymbol{C}, \\ \boldsymbol{E}\,\mathcal{Q}\,\boldsymbol{A}^\top + \boldsymbol{A}\,\mathcal{Q}\,\boldsymbol{E} &= -\boldsymbol{B}\,\boldsymbol{B}^\top. \end{aligned} \tag{43}$$

If system (42) is asymptotically stable, i.e., all eigenvalues of the matrix pencil $\lambda\boldsymbol{E}-\boldsymbol{A}$ have negative real part, then the Lyapunov equations (43) have unique solutions $\mathcal{P}$ and $\mathcal{Q}$ that are symmetric and positive semidefinite. If system (42) is unstable, we can define the Gramians as solutions of the projected Lyapunov equations as it has been done for differential-algebraic equations [34].

The Gramians $\mathcal{P}$ and $\mathcal{Q}$ can be used to compute the minimal input energy $\mathcal{E}_\eta = \|\eta\|^2_{L^2(-\infty,0)}$ that is needed to reach a state ξ_0 at $t = 0$ from a zero state at $t = -\infty$ and also the resulting output energy $\mathcal{E}_\varrho = \|\varrho\|^2_{L^2(0,\infty)}$ of (42) with the initial state $\xi(0) = \xi_0$ and $\eta(t) = 0$ for $t \geq 0$. If system (42) is controllable, i.e., $\operatorname{rank}[\lambda\boldsymbol{E} - \boldsymbol{A}^\top,\, \boldsymbol{C}^\top] = n$ for all complex λ, then $\mathcal{P}$ is nonsingular and we have

$$\min_{\eta\in L^2(-\infty,0)} \mathcal{E}_\eta = \xi_0^\top\mathcal{P}^{-1}\xi_0, \qquad\qquad \mathcal{E}_\varrho = \xi_0^\top\boldsymbol{E}\,\mathcal{Q}\,\boldsymbol{E}\,\xi_0. \tag{44}$$

The first relation implies that a large amount of the input energy $\mathcal{E}_\eta$ is required to reach the state ξ_0 which lies in an invariant subspace of $\mathcal{P}$ corresponding to its small eigenvalues. Such a state is difficult to reach. On the other hand, it follows from the second relation in (44) that if ξ_0 is contained in an invariant subspace of $\boldsymbol{E}\,\mathcal{Q}\,\boldsymbol{E}$ corresponding to its small eigenvalues, then the initial state $\xi(0) = \xi_0$ has a small effect on the output energy $\mathcal{E}_\varrho$ and it is difficult to observe. System (42) is *balanced* if $\mathcal{P} = \mathcal{Q} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. The diagonal elements σ_j are called the *Hankel singular values*. Clearly, the states of a balanced system related to the small Hankel singular values are difficult to reach and to observe at the same time. The truncation of such states essentially does not change the input-output relation of the system. Note that the Hankel singular values of (42) are invariant under state space transformation and they can be computed as the classical singular values of a matrix $L^\top\boldsymbol{E}R$, where R and L are the Cholesky factors of the Gramians $\mathcal{P} = RR^\top$ and $\mathcal{Q} = LL^\top$.

The balanced truncation model reduction approach consists in a transformation of system (42) into a balanced form and a truncation of the states that correspond to the small Hankel singular values. In practice, we do not actually need to compute the balancing transformation explicitly. Instead, we can combine balancing and truncation by performing the projection

$$\tilde{\boldsymbol{E}} = \boldsymbol{W}^\top\boldsymbol{E}\boldsymbol{S}, \qquad \tilde{\boldsymbol{A}} = \boldsymbol{W}^\top\boldsymbol{A}\boldsymbol{S}, \qquad \tilde{\boldsymbol{B}} = \boldsymbol{S}^\top\boldsymbol{B}, \qquad \tilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{W}, \tag{45}$$

where the projection matrices $\boldsymbol{W}$ and $\boldsymbol{S}$ determine left and right subspaces corresponding to the dominant Hankel singular values of system (42). These matrices can be computed by the following algorithm.

Algorithm 3.1 Projection matrices for balanced truncation
Given $\boldsymbol{\mathcal{T}} = [\,\boldsymbol{E},\, \boldsymbol{A}^\top,\, \boldsymbol{C}^\top,\, \boldsymbol{B}^\top\,]$, compute the projection matrices $\boldsymbol{W}$ and $\boldsymbol{S}$.

1. Compute the Cholesky factors R and L of the controllability and observability Gramians $\mathcal{P} = RR^\top$ and $\mathcal{Q} = LL^\top$.
2. Compute the singular value decomposition

$$L^\top \boldsymbol{E} R = [\,U_1,\, U_2\,] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} [\,V_1,\, V_2\,]^\top,$$

   where the matrices $[\,U_1,\, U_2\,]$ and $[\,V_1,\, V_2\,]$ have orthonormal columns,

$$\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_\ell), \quad \Sigma_2 = \operatorname{diag}(\sigma_{\ell+1}, \ldots, \sigma_r),$$

   with $\sigma_1 \geq \ldots \geq \sigma_\ell \gg \sigma_{\ell+1} \geq \ldots \geq \sigma_r > 0$ and $r = \operatorname{rank}(L^\top \boldsymbol{E} R)$.
3. Compute the projection matrices $\boldsymbol{W} = LU_1\Sigma_1^{-1/2}$ and $\boldsymbol{S} = RV_1\Sigma_1^{-1/2}$.

The reduced-order system

$$\begin{aligned} \tilde{\boldsymbol{E}}\,\dot{\tilde{\xi}}(t) &= \tilde{\boldsymbol{A}}^\top \tilde{\xi}(t) + \tilde{\boldsymbol{C}}^\top \eta(t), \\ \tilde{\varrho}(t) &= \tilde{\boldsymbol{B}}^\top \tilde{\xi}(t) \end{aligned} \tag{46}$$

with the system matrices as in (45) has the transfer function

$$\tilde{\boldsymbol{\mathcal{T}}}(s) = \tilde{\boldsymbol{B}}^\top (s\tilde{\boldsymbol{E}} - \tilde{\boldsymbol{A}}^\top)^{-1} \tilde{\boldsymbol{C}}^\top.$$

One can show that (46) is asymptotically stable and the absolute error

$$\|\tilde{\varrho} - \varrho\|_{L^2(0,\infty)} \leq \|\tilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\|_\infty \|\eta\|_{L^2(0,\infty)} \tag{47}$$

holds, where

$$\|\tilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\|_\infty := \sup_{\omega \in \mathbb{R}} \|\tilde{\boldsymbol{\mathcal{T}}}(i\omega) - \boldsymbol{\mathcal{T}}(i\omega)\| \leq 2(\sigma_{\ell+1} + \ldots + \sigma_r). \tag{48}$$

This bound allows an adaptive choice of the state space dimension ℓ of the reduced model depending on how accurate the approximation is needed. We will use this fact in the following.

A main difficulty in balanced truncation model reduction for large-scale systems is that the matrix Lyapunov equations (43) have to be solved. However, recent results on low-rank approximations to the solutions of Lyapunov equations [3,21,25] make the balanced truncation model reduction approach viable for large-scale problems.

3.2 *Convergence analysis*

In this subsection we present the balanced truncation descent algorithm for the semi-discretized optimal control problem (12), (13) and investigate sufficient conditions for convergence. Again, the main idea of the approach consists in replacing the semi-discretized adjoint equation (27) by the approximate reduced-order system (40).

Algorithm 3.2 Balanced truncation descent algorithm (BTDM)

1. Consider an initial control $\boldsymbol{u}_0$.
2. Compute the projection matrices $\boldsymbol{W}$ and $\boldsymbol{S}$ applying Algorithm 3.1 to the system $\boldsymbol{\mathcal{T}} = [\,\boldsymbol{E},\, \boldsymbol{A}^\top,\, \boldsymbol{C}^\top,\, \boldsymbol{B}^\top\,]$.
3. FOR $k = 0, 1, \ldots$
   (a) solve the semi-discretized state equation

$$\boldsymbol{E}\dot{\boldsymbol{y}}_k = \boldsymbol{A}\boldsymbol{y}_k + \boldsymbol{N}(\boldsymbol{y}_k) + \boldsymbol{B}\boldsymbol{u}_k, \quad \boldsymbol{y}_k(0) = \boldsymbol{g};$$

   (b) compute $\tilde{\boldsymbol{p}}_k$ as the solution of the reduced-order semi-discretized adjoint equation

$$-\tilde{\boldsymbol{E}}\dot{\tilde{\boldsymbol{p}}}_k = \big(\tilde{\boldsymbol{A}}^\top + \tilde{\boldsymbol{F}}(\boldsymbol{y}_k)\big)\tilde{\boldsymbol{p}}_k + \tilde{\boldsymbol{C}}^\top\boldsymbol{Q}(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{y}_k), \quad \tilde{\boldsymbol{p}}_k(T) = 0$$

   with $\tilde{\boldsymbol{E}} = \boldsymbol{W}^\top\boldsymbol{E}\boldsymbol{S}$, $\tilde{\boldsymbol{A}} = \boldsymbol{S}^\top\boldsymbol{A}\boldsymbol{W}$, $\tilde{\boldsymbol{F}}(\boldsymbol{y}_k) = \boldsymbol{W}^\top\boldsymbol{F}(\boldsymbol{y}_k)\boldsymbol{S}$, $\tilde{\boldsymbol{B}} = \boldsymbol{S}^\top\boldsymbol{B}$ and $\tilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{W}$;
   (c) update $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \beta_k H_k(\tilde{\boldsymbol{B}}^\top\tilde{\boldsymbol{p}}_k - \alpha\boldsymbol{R}\boldsymbol{u}_k)$, with H_k positive definite.
   END FOR

The following lemma gives a bound on the error $\tilde{\boldsymbol{B}}^\top\tilde{\boldsymbol{p}} - \boldsymbol{B}^\top\boldsymbol{p}$, where $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$ satisfy (27) and (40), respectively.

LEMMA 3.3 *Let $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$ be the solutions of the adjoint equation* (27) *and the reduced-order system* (40), (41), *where the projection matrices $\boldsymbol{W}$ and $\boldsymbol{S}$ are computed by Algorithm* 3.1 *applied to $\boldsymbol{\mathcal{T}} = [\,\boldsymbol{E},\, \boldsymbol{A}^\top,\, \boldsymbol{C}^\top,\, \boldsymbol{B}^\top\,]$. Let $\delta = 2(\sigma_{\ell+1} + \ldots + \sigma_r)$, where $\sigma_{\ell+1}, \ldots, \sigma_r$ are the truncated Hankel singular*

values of (27)*. Then*

$$\|\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}} - \boldsymbol{B}^\top \boldsymbol{p}\|_{L^2(0,T)} \le \big(\delta\|\boldsymbol{Q}\| + \kappa_5(\boldsymbol{y}) + \tilde{\kappa}_5(\boldsymbol{y})\big)\, \|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|_{L^2(0,T)}\,, \tag{49}$$

where $\kappa_5(\boldsymbol{y})$ *and* $\tilde{\kappa}_5(\boldsymbol{y})$ *are some constants depending on* $\boldsymbol{y}$*.*

Proof Let ζ and $\tilde{\zeta}$ be the solutions of the systems

$$-\boldsymbol{E}\dot{\zeta} = \boldsymbol{A}^\top \zeta + \boldsymbol{C}^\top \eta, \quad \zeta(T) = 0, \tag{50}$$

and

$$-\tilde{\boldsymbol{E}}\dot{\tilde{\zeta}} = \tilde{\boldsymbol{A}}^\top \tilde{\zeta} + \tilde{\boldsymbol{C}}^\top \eta, \quad \tilde{\zeta}(T) = 0,$$

respectively, where the input η is given by

$$\eta(t) = \begin{cases} \boldsymbol{Q}(\boldsymbol{z}(t) - \boldsymbol{C}\boldsymbol{y}(t)), & 0 \le t \le T, \\ 0, & t > T. \end{cases}$$

Clearly, $\zeta(t) = \xi(T-t)$ and $\tilde{\zeta}(t) = \tilde{\xi}(T-t)$, where ξ and $\tilde{\xi}$ satisfy (42) and (46), respectively. Then we have

$$\begin{aligned} \|\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}} - \boldsymbol{B}^\top \boldsymbol{p}\|_{L^2(0,T)} \le{}& \|\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}} - \tilde{\boldsymbol{B}}^\top \tilde{\zeta}\,\|_{L^2(0,T)} + \|\tilde{\boldsymbol{B}}^\top \tilde{\zeta} - \boldsymbol{B}^\top \zeta\,\|_{L^2(0,T)} \\ &+ \|\boldsymbol{B}^\top \zeta - \boldsymbol{B}^\top \boldsymbol{p}\|_{L^2(0,T)}. \end{aligned}$$

Using the error bounds (47) and (48) we can estimate

$$\begin{aligned} \|\tilde{\boldsymbol{B}}^\top \tilde{\zeta} - \boldsymbol{B}^\top \zeta\,\|_{L^2(0,T)} = {}& \|\tilde{\boldsymbol{B}}^\top \tilde{\xi} - \boldsymbol{B}^\top \xi\,\|_{L^2(0,T)} \le \delta\,\|\,\eta\,\|_{L^2(0,\infty)} \\ \le{}& \delta\,\|\boldsymbol{Q}\|\,\|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|_{L^2(0,T)}. \end{aligned} \tag{51}$$

Consider now the vector $\boldsymbol{p}(t) - \zeta(t)$ that satisfies the system

$$\begin{aligned} -\boldsymbol{E}(\dot{\boldsymbol{p}}(t) - \dot{\zeta}(t)) &= \boldsymbol{A}^\top(\boldsymbol{p}(t) - \zeta(t)) + \boldsymbol{F}(\boldsymbol{y}(t))\boldsymbol{p}(t), \\ \boldsymbol{p}(T) - \zeta(T) &= 0. \end{aligned}$$

This system can be rewritten in integral form as

$$\boldsymbol{p}(t) - \zeta(t) = \boldsymbol{E}^{-1}\boldsymbol{A}^\top \int_t^T (\boldsymbol{p}(\tau) - \zeta(\tau))\, d\tau + \boldsymbol{E}^{-1} \int_t^T \boldsymbol{F}(\boldsymbol{y}(\tau))\boldsymbol{p}(\tau)\, d\tau.$$

Since $\boldsymbol{F}(\boldsymbol{y}) = \boldsymbol{N}'(\boldsymbol{y})^\top$, we obtain that

$$\begin{aligned}\|\boldsymbol{p}(t) - \zeta(t)\| &\le \|\boldsymbol{E}^{-1}\|\Big(\|\boldsymbol{A}\| \int_t^T \|\boldsymbol{p}(\tau) - \zeta(\tau)\|\, d\tau + \int_t^T \|\boldsymbol{F}(\boldsymbol{y}(\tau))\|\|\boldsymbol{p}(\tau)\|\, d\tau\Big)\\ &\le \|\boldsymbol{E}^{-1}\|\Big(\|\boldsymbol{A}\| \int_t^T \|\boldsymbol{p}(\tau) - \zeta(\tau)\|\, d\tau + \|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}\|\boldsymbol{p}\|_{L^2(0,T)}\Big).\end{aligned}$$

Using Gronwall's inequality [28], we get

$$\|\boldsymbol{p}(t) - \zeta(t)\| \le \|\boldsymbol{E}^{-1}\|\|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}\|\boldsymbol{p}\|_{L^2(0,T)} e^{\|\boldsymbol{E}^{-1}\|\|\boldsymbol{A}\|(T-t)}$$

and, hence,

$$\begin{aligned}\|\boldsymbol{p} - \zeta\|_{L^2(0,T)} &\le \|\boldsymbol{E}^{-1}\|\|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}\|\boldsymbol{p}\|_{L^2(0,T)} \left(\int_0^T e^{2\|\boldsymbol{E}^{-1}\|\|\boldsymbol{A}\|(T-\tau)}\, d\tau\right)^{1/2}\\ &= \|\boldsymbol{E}^{-1}\|\|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}\|\boldsymbol{p}\|_{L^2(0,T)} \left(\frac{e^{2T\|\boldsymbol{E}^{-1}\|\|\boldsymbol{A}\|} - 1}{2\|\boldsymbol{E}^{-1}\|\|\boldsymbol{A}\|}\right)^{1/2}\\ &= \kappa_3(\boldsymbol{y})\|\boldsymbol{p}\|_{L^2(0,T)},\end{aligned}$$

where $\kappa_3(\boldsymbol{y}) = \sqrt{\|\boldsymbol{E}^{-1}\|(e^{2T\|\boldsymbol{E}^{-1}\|\|\boldsymbol{A}\|} - 1)/(2\|\boldsymbol{A}\|)}\, \|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}$.

Using again Gronwall's inequality, we obtain from equation (27) the following estimate

$$\|\boldsymbol{p}\|_{L^2(0,T)} \le \kappa_4(\boldsymbol{y})\|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|_{L^2(0,T)},$$

where

$$\kappa_4(\boldsymbol{y}) = \|\boldsymbol{C}\|\|\boldsymbol{Q}\|\sqrt{\frac{\|\boldsymbol{E}^{-1}\|(e^{2T\|\boldsymbol{E}^{-1}\|(\|\boldsymbol{A}\| + \|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)})} - 1)}{2\|\boldsymbol{A}\| + 2\|\boldsymbol{F}(\boldsymbol{y})\|_{L^2(0,T)}}}.$$

Thus,

$$\|\boldsymbol{p} - \zeta\|_{L^2(0,T)} \le \kappa_5(\boldsymbol{y})\|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|_{L^2(0,T)} \tag{52}$$

with $\kappa_5(\boldsymbol{y}) = \kappa_3(\boldsymbol{y})\kappa_4(\boldsymbol{y})$. Analogously, we get

$$\|\tilde{\boldsymbol{p}} - \tilde{\zeta}\|_{L^2(0,T)} \le \tilde{\kappa}_5(\boldsymbol{y})\|\boldsymbol{C}\boldsymbol{y} - \boldsymbol{z}\|_{L^2(0,T)}, \tag{53}$$

where $\tilde{\kappa}_5(\boldsymbol{y})$ is as $\kappa_5(\boldsymbol{y})$ with $\boldsymbol{E}$, $\boldsymbol{A}$ and $\boldsymbol{F}$ replaced by $\tilde{\boldsymbol{E}}$, $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{F}}$, respectively.

Finally, combining (51), (52) and (53) we obtain estimate (49). □

THEOREM 3.4 *Let H_k be positive definite with coercivity constant c_k. If for an appropriate chosen δ, the condition*

$$(\delta \|\boldsymbol{Q}\| + \kappa_5(\boldsymbol{y}_k) + \tilde{\kappa}_5(\boldsymbol{y}_k)) \|\boldsymbol{C}\boldsymbol{y}_k - \boldsymbol{z}\|_{L^2(0,T)} \leq \frac{c_k(1-\lambda)}{c_k + \|H_k\|} \left\|\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\right\|_{L^2(0,T)} \tag{54}$$

holds for some given $0 < \lambda < 1$, then the direction $\tilde{\boldsymbol{d}}_k = -H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k)$ is a descent direction at $\boldsymbol{u}_k$.

Proof Multiplying the exact gradient by $-\tilde{\boldsymbol{d}}_k$ we obtain that

$$\begin{aligned}
&(\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k), H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k))_{L^2(0,T)} \\
&\quad = (\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k, H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k))_{L^2(0,T)} \\
&\quad = (\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k, H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k + \boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k))_{L^2(0,T)} \\
&\quad \geq c_k \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|^2_{L^2(0,T)} + \left(\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k, H_k(\boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k)\right)_{L^2(0,T)}.
\end{aligned}$$

From the Cauchy-Schwarz inequality it then follows that

$$\begin{aligned}
&(\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k), H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k))_{L^2(0,T)} \\
&\qquad \geq \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|_{L^2(0,T)} \Big(c_k \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|_{L^2(0,T)} \\
&\qquad\qquad - \|H_k\| \|\boldsymbol{B}\boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k)\|_{L^2(0,T)}\Big) \qquad (55) \\
&\qquad \geq \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|_{L^2(0,T)} \Big(c_k \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)} \\
&\qquad\qquad - (c_k + \|H_k\|) \|\boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)}\Big). \qquad (56)
\end{aligned}$$

Using estimate (49) and hypothesis (54) we obtain that

$$\begin{aligned}(c_k + \|H_k\|)\|\boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)} \\ &\leq (c_k + \|H_k\|)(\delta \|\boldsymbol{Q}\| + \kappa_5(\boldsymbol{y}_k) + \tilde{\kappa}_5(\boldsymbol{y}_k)) \|\boldsymbol{C}\boldsymbol{y}_k - \boldsymbol{z}\|_{L^2(0,T)} \\ &\leq c_k(1-\lambda)\|\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)} \\ &\leq c_k\|\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)} - \rho_k \|H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k)\|_{L^2(0,T)}\end{aligned}$$

with $\rho_k = \lambda \frac{c_k}{\|H_k\|}$. Therefore, (56) implies that

$$\begin{aligned}(\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k),\ H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k))_{L^2(0,T)} \\ &\geq \rho_k \|\boldsymbol{\mathcal{J}}'(\boldsymbol{u}_k)\|_{L^2(0,T)} \|H_k(\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k)\|_{L^2(0,T)}.\end{aligned}$$

□

Remark 3.5 It follows from (55) that

$$\begin{aligned}c_k\|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|_{L^2(0,T)} - \|H_k\|\|\boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)} \\ &\geq \vartheta \|\alpha \boldsymbol{R}\boldsymbol{u}_k - \boldsymbol{B}^\top \boldsymbol{p}_k\|_{L^2(0,T)}, \quad (57)\end{aligned}$$

for some constant $\vartheta > 0$, directly implies that $\tilde{\boldsymbol{d}}_k$ is a descent direction. This condition is numerically verified in the experiments below. In practice, however, there is no adjoint state information available and therefore this condition cannot be verified. Differently from (57), condition (54) can be verified before the application of the BTDM, by evaluating the different constants involved.

4 Numerical results

In this section we present some numerical experiments, which show the main features of the BTDM described in Algorithm 3.2. By means of two- and one-dimensional example problems, the dependence of this numerical approach on different parameter values such as viscosity coefficient, control weight or mesh size are investigated and a comparison with the BFGS and the gradient methods is carried out.

For the solution of the partial differential equations involved, a homogeneous finite differences scheme for the space discretization is used in both examples. For the numerical solution of the semi-discretized systems an implicit Euler method is applied. The state equation is fully solved with the nonlinear term

involving information from the previous time step. Unless otherwise specified, the space step $h = 0.01$ and the time step $\tau = 0.01$ were considered.

The methods stop if the difference norm of two consecutive iterates reaches a precision value ε. The methods begin with a control initialization value $\boldsymbol{u}_0 \equiv 0$. For simplicity, we choose the parameter β to be constant and equal to 0.5 in each descent iteration. It is also possible to use alternative line search strategies [12], but we have not considered this, since the computational cost of such strategies appears to be high.

Unless otherwise specified, the numerical experiments were carried out using MATLAB. For solving the Lyapunov matrix equations we used the LYAPACK Toolbox [26].

4.1 *Semilinear system*

In this first example we consider the following optimal control problem

$$\begin{array}{ll} \text{minimize} & J(y,u) = \dfrac{1}{2}\displaystyle\int_0^T \int_\Omega \|\mathcal{C}y - z\|^2 \, dx\, dt + \dfrac{\alpha}{2}\int_0^T \int_\Omega \|\mathcal{B}u\|^2 \, dx\, dt \\ \text{subject to} & \dfrac{\partial y}{\partial t} - \Delta y + \theta y^3 = \mathcal{B}u \quad \text{in } (0,T)\times\Omega, \\ & y(t,x) = 0 \quad \text{in } (0,T)\times\Gamma, \\ & y(0,x) = 0.01 x_1 x_2 \quad \text{in } \Omega, \end{array}$$

where $\Omega = (0,1)\times(0,1)$, Γ is the boundary of Ω and $\theta > 0$. As desired state we choose the function $z = x_1 x_2$. We consider as the control and observation domains the five central points of the spatial mesh. The order of the semi-discretized full order adjoint system is $n = 9801$. It has been approximated by a reduced model of order $\ell = 5$ with error bound $\delta = 7.268 \times 10^{-5}$.

The cubic nonlinearity together with the parameter θ are responsible for a monotonic behaviour of the evolution operator. In Table 1 we present the number of iterations and the quotient of computing times for different values of θ. One can see that the total computing time for the BTDM is less than half of the computing time needed by the standard BFGS method. This is certainly not unexpected, since the adjoint system is reduced, but not the state equation. It can also be noted from the data that the number of the BTDM iterations are similar to those of the BFGS method and do not differ as θ increases.

In Table 2, the data for different values of α are reported. The convergence of the BFGS and BTDM algorithms needs more iterations and time as α decreases. The behaviour in both cases with respect to iteration number is, however, similar.

Next, the behaviour of the BTDM for different mesh sizes h is investigated.

Table 1. Semilinear system ($h = 0.01$, $\tau = 0.01$, $\alpha = 0.1$, $z = x_1x_2$, $\varepsilon = 10^{-4}$): number of iterations and computing time for different values of θ.

θ	iter. BFGS	iter. BTDM	time BTDM/ time BFGS
1	5	4	0.3806
10	5	4	0.3816
100	5	4	0.3785
1000	5	4	0.4585
10000	5	4	0.3824

Table 2. Semilinear system ($h = 0.01$, $\tau = 0.01$, $\theta = 1$, $z = x_1x_2$, $\varepsilon = 10^{-4}$): number of iterations and computing time for different values of α.

α	iter. BFGS	iter. BTDM	time BTDM/ time BFGS
1	3	3	0.5998
0.1	5	4	0.3806
0.01	5	5	0.4880
0.001	5	5	0.4864

Table 3. Semilinear system ($\tau = 0.01$, $\theta = 1$, $\alpha = 0.5$, $z = x_1x_2$, $\varepsilon = 10^{-5}$): number of iterations and computing time for different mesh sizes h.

1/h	iter. BFGS	iter. BTDM	time BTDM /time BFGS
40	14	14	0.5857
48	14	14	0.5126
56	15	15	0.6044
64	15	15	0.5074

The observation and control domains utilized in this case are depicted in Figure 1. We look for a control $u(t,x) = \chi_{\Omega_c}(x)\phi(t)$, where $\phi : [0,T] \mapsto \mathbb{R}$ and χ_{Ω_c} is the indicator function of the control domain Ω_c. The observation vector consists of all points included in the sector Ω_o. In Table 3, the convergence data for the BFGS method and the BTDM for different mesh sizes are given. The number of iterations of both methods does not differ significantly as the mesh step becomes smaller.

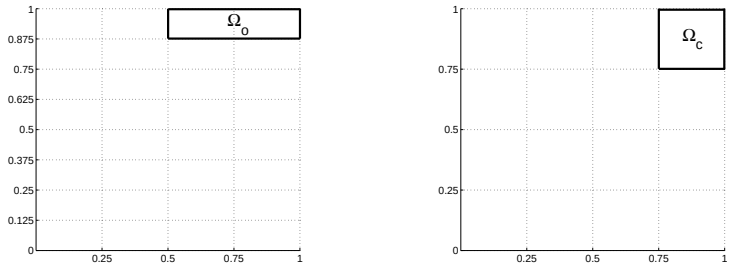

Figure 1. Observation and control domains.

Finally, an alternative time dependent desired state was considered. The control problem consisted in tracking the function $z = (x_1x_2 - 1)\sin(10t)$ with

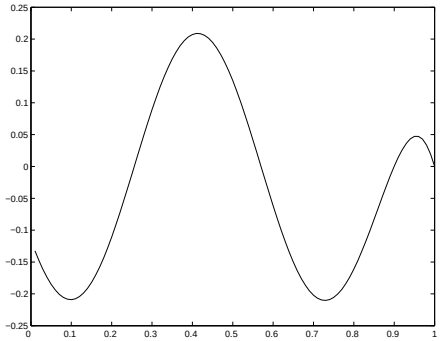

Figure 2. Optimal control function $\phi^*(t)$.

the separated control $u(t,x) = \chi_\Omega(x)\phi(t)$. The observation domain is given by the five central points of the mesh. The optimal control function $\phi^*(t)$ is depicted in Figure 2. In this case, both the BFGS and BTDM methods needed 14 iterations to converge with a precision $\varepsilon = 10^{-5}$. The remaining parameters took the values $\alpha = 0.01$, $\tau = 0.01$ and $h = 1/24$. Despite of the difficulties related to the active tracking of the time-dependent target, the BTDM behaves efficiently in this case.

4.2 *Burgers equation*

In this example we consider the optimal control of the instationary Burgers equation. This equation is known to be a good one-dimensional model for turbulence and posses important features of fluid flow phenomena. Specifically, we consider the following optimal control problem

$$\begin{array}{ll} \text{minimize} & J(y,u) = \dfrac{1}{2}\displaystyle\int_0^T\int_0^1 |\mathcal{C}y - z|^2 \, dx\, dt + \dfrac{\alpha}{2}\int_0^T\int_0^1 |\mathcal{B}u|^2 \, dx\, dt \\ \text{subject to} & \dfrac{\partial y}{\partial t} - \nu\Delta y + y'y = \mathcal{B}u, \\ & y(t,0) = 0, \;\; y(t,1) = 0, \\ & y(0,x) = \sin(4\pi x), \end{array}$$

where ν is the viscosity coefficient and y' is the spatial partial derivative of y. To fit in our framework, we consider the nonlinear operator $\mathcal{N}(v) = \mathcal{N}_2(v,v)$, where $\mathcal{N}_2(v,w) = \frac{1}{3}[(vw)' + vw']$ satisfies conditions (4).

As control and observation domains we consider first the three central points of the mesh. The target is to reach the stationary desired state $z = 0$ in a time horizon $T = 0.1$. The remaining parameter data are $\nu = 1/800$ and $\alpha = 0.1$. The semi-discretized full order adjoint system is of order $n = 199$.

With a stopping parameter $\varepsilon = 0.0005$, the gradient method takes 30 itera-

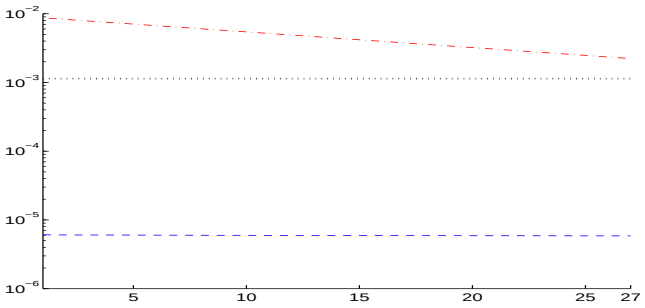

Figure 3. Convergence condition: $\left\|\alpha \boldsymbol{R}\boldsymbol{u}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\right\|_{L^2(0,T)}$ $(\cdot -)$, $(\delta \|\boldsymbol{Q}\| + \kappa_5(\boldsymbol{y}_k) + \tilde{\kappa}_5(\boldsymbol{y}_k)) \|\boldsymbol{C}\boldsymbol{y}_k - \boldsymbol{z}\|_{L^2(0,T)}$ $(\cdot\cdot)$ and $\|\boldsymbol{B}^\top \boldsymbol{p}_k - \tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{p}}_k\|_{L^2(0,T)}(-)$

tions to converge, while the related BTDM algorithm with $H_k = I$ stops after 27 iterations.

In Figure 3, the verification of the descent sufficient conditions (54) and (57) is carried out. The numerical values of the adjoint error, the left hand side term in (54) and the descent direction norm are plotted. For the evaluation of the integrals we used a trapezoidal rule, while the matrix 2-norms were estimated by using the inequality

$$\|\cdot\| \leq \sqrt{\|\cdot\|_1 \|\cdot\|_\infty}.$$

The satisfaction of the descent sufficient conditions (54) and (57), see Remark 3.5, of the BTDM can be inferred from the plot. Comparing the adjoint error norm and the descent direction norm, always a gap, which implies the convergence of our method, occurs. However, also a gap between both sufficient conditions exists. This fact provokes that in many numerical examples the satisfaction of (57) can be observed, and therefore the convergence of BTDM, but (54) can be numerically verified only for the first iterations. This fact suggests that the result of Theorem 3.4 may be improved by using more precise estimates.

In Table 4, the number of iterations for the descent method (DM) and the BTDM algorithm is given for different viscosity coefficient values, the time horizon $T = 1$ and the initial condition $y(0, x) = \sin(4\pi x)$. The control and final control state for $\nu = 1/200$ are plotted in Figure 4. Although no monotonic behavior of the methods with respect to the coefficient can be observed from the data, the number of iterations for both methods does not differ significantly and the BTDM behaves robustly.

Table 4. Example 2: $h = 0.005$, $\tau = 0.01$, $\alpha = 0.1$, $z = 0$, $\varepsilon = 0.0005$.

ν	iter. DM	iter. BTDM
1/10	55	55
1/50	79	79
1/100	53	53
1/150	40	38
1/200	41	38

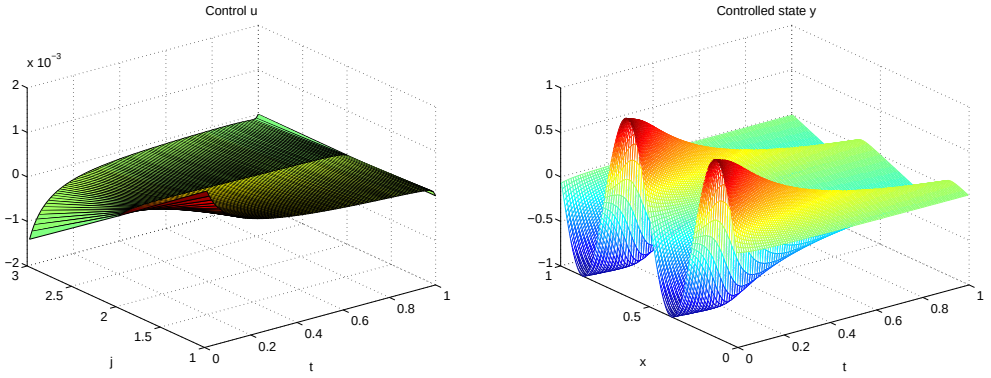

Figure 4. Final control and controlled state; $\nu = 1/200$.

5 Conclusions

According to the analysis and the numerical experiments carried out, we conclude the following:

- The BTDM adjoint evaluation presented in this paper is an alternative approach to obtain gradient related information in optimal control problems for nonlinear evolution PDEs. From the balanced truncation estimates, a convergence analysis of the reduced descent method can be carried out. As numerical experiments show, the convergence condition given in Theorem 3.4 seems not to be restrictive in practice.

- The BTDM behaves similar to the correspondent descent method with respect to the number of iterations. Since the same projection matrices are used in each BTDM iteration and only the reduced-order adjoint equation is solved, the computing time for the BTDM is less than the time needed for the descent method.

References

[1] Afanasiev, K. and Hinze, M., 2001, Adaptive control of a wake flow using proper orthogonal decomposition. In: J. Cagnol, M. P. Polis and J.-P. Zolesio (Eds) *Shape Optimization and Optimal Design*, Lecture Notes in Pure and Applied Mathematics 216 (Marcel Dekker), pp. 317-332.

[2] Alekseev V., Tikhomirov V., Fomin S., 1982, *Commande Optimale* (Moskow: MIR).
[3] Baur, U. and Benner, P., 2004, Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic. Preprint 161, DFG Research Center MATHEON, TU Berlin.
[4] Benner, P., Mehrmann, V. and Sorensen, D. (Eds), 2005, *Dimension Reduction of Large-Scale Systems*, Lecture Notes in Computational Science and Engineering 45 (Berlin/Heidelberg: Springer-Verlag).
[5] Benner, P., Quintana-Ortí, E.S. and Quintana-Ortí, G., 2003, State-space truncation methods for parallel model reduction of large- scale systems. *Parallel Comput.*, **29**, 1701–1722.
[6] Dautray, R. and Lions J.L., 2000, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 5 (Berlin: Springer-Verlag).
[7] Glover, K., 1984, All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *Internat. Control*, **39**, 1115–1193.
[8] Griesse, R. and Walther, A., 2004, Evaluating gradients in optimal control: continuous adjoints versus automatic differentiation. *J. Optim. Theory Appl.*, **122**, 63–86.
[9] Hinze, M. and Slawig, T., 2003, Adjoint gradients compared to gradients from algorithmic differentiation in instantaneous control of the Navier-Stokes equations. *Optim. Methods Software*, **18**, 299–315.
[10] Hinze, M. and Volkwein, S., 2005, Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control. In: P. Benner, V. Mehrmann and D. Sorensen (Eds) *Dimension Reduction of Large-Scale Systems*, Lecture Notes in Computational Science and Engineering 45 (Berlin/Heidelberg: Springer-Verlag), 263–308.
[11] Hinze, M. and Pinnau, R. and Ulbrich, M. and Ulbrich, S., 2009, *Optimization with PDE constraints*, (Berlin: Springer- Verlag).
[12] Kelley, C.T., 1999, *Iterative methods for optimization*, Frontiers in Applied Mathematics, Vol.18 (Philadelphia: SIAM).
[13] Kelley, C.T. and Sachs, E., 1987, Quasi-Newton methods and unconstrained optimal control problems. *SIAM J. Control Optim.*, **25**, 1503–1516.
[14] Kelley, C.T. and Sachs, E., 1991, A new proof of superlinear convergence for Broyden's method in Hilbert space. *SIAM J. Optim.*, **1**, 146–150.
[15] Kunisch, K.,*Nonlinear optimization in infinite dimensions*, Lecture Notes, TU Berlin.
[16] Kunisch, K. and Volkwein, S., 2002, Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics, *SIAM J. Numer. Anal.*, **40**, 492–515.
[17] Kupfer, F.-S., 1996, An infinite-dimensional convergence theory for reduced SQP methods in Hilbert space, *SIAM J. Optim.*, Vol. 6, 126–163.
[18] Krasnoselskii, M. A. and Zabreiko, P. P. and Pustylnik, E. I. and Sobolevskii, P. E., *Integral operators in spaces of summable functions*, Noordhoff International Publishing, Leiden, 1976.
[19] Lall, S., Marsden, J. E. and Glavaški, S., 2002, A subspace approach to balanced truncation for model reduction of nonlinear control systems. *Internat. J. Robust Nonlinear Control*, **12**, 519–535.
[20] Laub, A. J., Heath, M.T., Paige C. C. and Ward, R.C., 1987, Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. *IEEE Trans. Automat. Control*, **32**, 115–122.
[21] Li, J.-R. and White, J., 2002, Low rank solution of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, **24**, 260–280.
[22] Liebermeister, W., Baur, U. and Klipp, E., 2005, Biochemical network models simplified by balanced truncation. *FEBS Journal*, **272**, 4034–4043.
[23] Luenberger, D., 1969, *Optimization by Vector Space Methods* (New York: John Wiley and Sons).
[24] Moore, B.C., 1981, Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Automat. Control*, **26**, 17–32.
[25] Penzl, T., 1999/2000, A cyclic low-rank Smith method for large sparse Lyapunov equations, *SIAM J. Sci. Comput.*, **21**, 1401–1418.
[26] Penzl, T., 2000, LYAPACK Users Guide, Preprint SFB 393/00-33, Fakultät für Mathematik, Technische Universität Chemnitz. Available online at: www.netlib.org/lyapack
[27] Polak, E., 1973, An historical survey of computational methods in optimal control, *SIAM Rev.*, Vol. 15, 553–584.
[28] Quarteroni, A. and Valli, A., 2005, *Numerical Approximation of Partial Differential Equations* (Berlin/Heidelberg: Springer-Verlag).
[29] Rewieński, M.J., 2003, A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems, PhD thesis, Massachusetts Institute of Technology.
[30] Rowley, C.W., 2005, Model reduction for fluids, using balanced proper orthogonal decomposition.

Internat. J. Bifur. Chaos Appl. Sci. Engrg., **15**, 997–1013.
[31] Sandberg, H. and Rantzer, A., 2004, Balanced truncation of linear time-varying systems. *IEEE Trans. Automat. Control*, **49**, 217–229.
[32] Scherpen, J.M.A., 1994, Balancing for nonlinear systems, PhD thesis, University of Twente.
[33] Shokoohi, S., Silverman, L.M. and Van Dooren, P., 1983, Linear time-variable systems: balancing and model reduction. *IEEE Trans. Automat. Control*, **28**, 810–822.
[34] Stykel, T., 2004, Gramian-Based Model Reduction for Descriptor Systems. *Math. Control Signals Systems*, **16**, 297–319.
[35] Stykel, T., 2006, Balanced truncation model reduction for semi-discretized Stokes equation. *Linear Algebra Appl.*, **415**, 262–289.
[36] Temam, R., 1988, *Infinite Dimensional Dynamical Systems in Mechanics and Physics* (New York: Springer-Verlag).
[37] Tröltzsch, F., 2005, *Optimale Steuerung partieller Differentialgleichungen* (Vieweg-Verlag).
[38] Verriest, E.I. and Kailath, T., 1983, On generalized balanced realizations. *IEEE Trans. Automat. Control*, **28**, 833–844.
[39] Willcox, K. and Peraire J., 2002, Balanced model reduction via the proper orthogonal decomposition. *AIAA Journal*, **40**, 2323–2330.